Naval Research Laboratory

Stennis Space Center, MS 39529-5004

NRL/MR/7440--08-9130

Visual Analysis of North Atlantic Hurricane Trends Using Parallel Coordinates and Statistical Techniques

Chad A. Steed

Mapping, Charting, and Geodesy Branch
Marine Geosciences Division

Patrick J. Fitzpatrick

Northern Gulf Institute, Mississippi State University
Stennis Space Center, Mississippi

T.J. Jankun-Kelly
J. Edward Swan II

Department of Computer Science and Engineering
Mississippi State University, Mississippi

July 7, 2008

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
07-07-2008

2. REPORT TYPE
Memorandum Report

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE

Visual Analysis of North Atlantic Hurricane Trends
Using Parallel Coordinates and Statistical Techniques

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Chad A. Steed, Patrick J. Fitzpatrick, T.J. Jankun-Kelly, and J. Edward Swan II

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
74-9531-08

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
Marine Geosciences Division
Stennis Space Center, MS 39529-5004

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/7440--08-9130

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
One Liberty Center
875 North Randolph St.
Arlington, VA 22203-1995

10. SPONSOR / MONITOR'S ACRONYM(S)
ONR

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The integration of automated statistical analysis capabilities with a highly interactive, multivariate visualization interface is presented in this paper. Innovative visual interaction techniques such as dynamic axis scaling, conjunctive parallel coordinates, statistical indicators, and aerial perspective shading are exploited to enhance the utility of classical parallel coordinate plots. Moreover, the system facilitates statistical processes such as stepwise regression and correlation analysis to assist in the identification and quantification of the most significant predictors for a particular dependent variable. These capabilities are combined into a unique visualization system that is demonstrated via a North Atlantic hurricane climate study using a systematic workflow. This research corroborates the notion that enhanced parallel coordinates coupled with statistical analysis can be used for more effective knowledge discovery and confirmation in complex, real-world data sets.

15. SUBJECT TERMS
Climate study; Multidimensional multivariate visualization; Correlation analysis; Visual interaction techniques; Focus+context; Statistical regression; Aerial perspective shading; Tropical cyclone; Axis scaling

16. SECURITY CLASSIFICATION OF:

a. REPORT
Unclassified

b. ABSTRACT
Unclassified

c. THIS PAGE
Unclassified

17. LIMITATION OF ABSTRACT
UL

18. NUMBER OF PAGES
21

19a. NAME OF RESPONSIBLE PERSON
Chad Steed

19b. TELEPHONE NUMBER *(include area code)*
(228) 688-4558

CONTENTS

1 Introduction

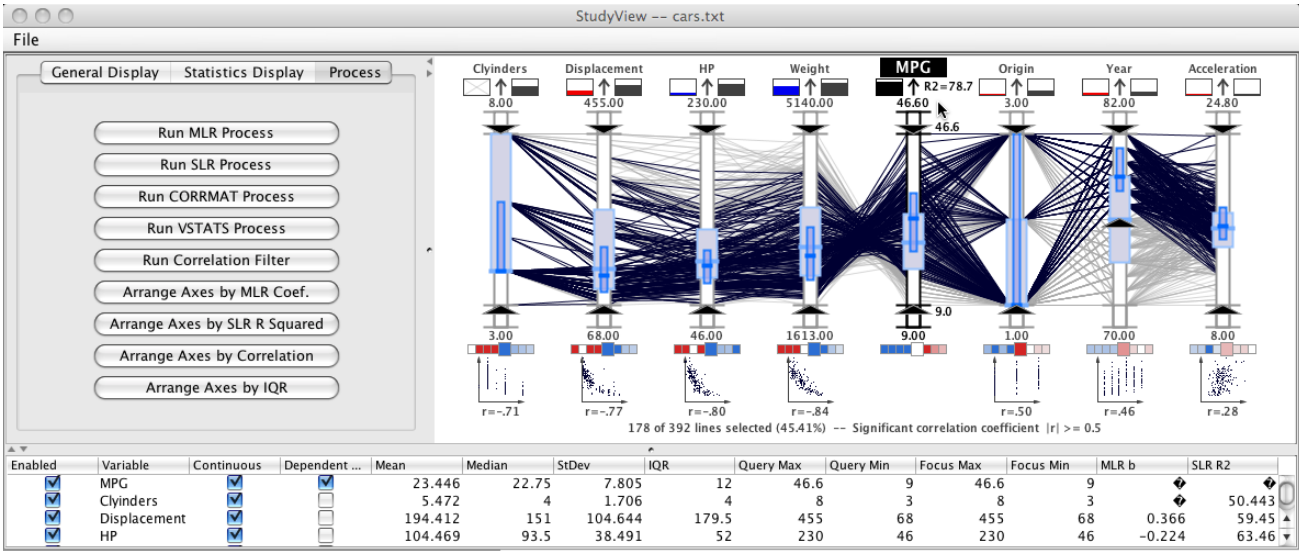
One of the most challenging tasks in multivariate data analysis is to identify and quantify the associations among sets of interrelated variables. In real-world climate studies, this task is even more daunting due to the uncertainty and complexity of dynamic, environmental data sets. Notwithstanding the difficulty, the variability and destructiveness of recent hurricane seasons has invigorated efforts by weather scientists to identify environmental variables that have the greatest impact on the intensity and frequency of seasonal hurricane activity. In general, the goal of such efforts is to improve the accuracy of seasonal forecasts which should, in turn, improve preparedness and reduce the impact of these devastating natural disasters.

One particularly useful method for predicting seasonal hurricane variability is based on the idea that there are predictors of the main dynamic parameters that affect storm activity, which can be observed up to a year in advance. Using historical data, the importance of these parameters is estimated using statistical regression techniques similar to those described by Vitart [1]. Although sometimes complicated to establish, these techniques provide an ordered list of the most important predictors for the dynamic parameters. Scientists gain additional insight in these studies by evaluating descriptive statistics and performing correlation analyses.

In conjunction with statistical analysis, researchers have relied on simple scatter plots and histograms which require several separate plots or layered plots to analyze multiple variables. Using separate plots, however, is not an optimal approach in this type of analysis due to perceptual issues such as change blindness (a phenomenon described by Rensink [2]), especially when searching for combinations of conditions. Although layered plots condense the information into a single display, there are issues due to occlusion and interference as demonstrated by Healey et al. [3]. Furthermore, the geographically-encoded data used in climate studies are usually displayed in the context of a geographical map; although certain important patterns (those directly related to geographic position) may be recognized in this context, additional information may be discovered more rapidly using non-geographical information visualization techniques. What's more, few multivariate visualization techniques provide access to integrated, automatic statistical analysis techniques commonly used in climate studies to identify significant associations. To compensate for these deficiencies, new visualization methods are needed that intelligently integrate statistical processes and accommodate the simultaneous display of real-world, multivariate data.

This paper discusses the extension and application of a popular multivariate information visualization technique, the parallel coordinate plot (PCP), to a hurricane climate study. The resulting system (see Fig. 1) provides a comprehensive environment for multivariate analysis by combining several innovative extensions to the classical PCP with automated statistical analyses. This paper also describes a systematic workflow for exploring environmental data with this system and concludes with a case study in which the system concepts are evaluated via climate analysis of seasonal intense hurricane activity. The results of this practical evaluation suggest that PCPs can be used in conjunction with statistical processes to more efficiently conduct real-world, multivariate data analysis on complex environmental data sets. Furthermore, this research effort fulfills the NIH/NSF Visualization Challenges Report recommendation that visualization researchers "collaborate closely with domain experts who have driving tasks in data-rich fields to produce tools and techniques that solve clear real-world needs [5]" through the inclusion of a hurricane expert throughout the design and evaluation of the system.

Manuscript approved June 16, 2008.



178 of 392 lines selected (45.41%) -- Significant correlation coefficient |r| >= 0.5

Enabled	Variable	Continuous	Dependent ...	Mean	Median	StDev	IQR	Query Max	Query Min	Focus Max	Focus Min	MLR b	SLR R2
☑	MPG	☑	☑	23.446	22.75	7.805	12	46.6	9	46.6	9	�	�
☑	Clyinders	☑	□	5.472	4	1.706	4	8	3	8	3	�	50.443
☑	Displacement	☑	□	194.412	151	104.644	179.5	455	68	455	68	0.366	59.45
☑	HP	☑	□	104.469	93.5	38.491	52	230	46	230	46	-0.224	63.46

Figure 1: The visualization system developed in this research is composed of a settings panel (upper left), parallel coordinates plot (upper right), and table view panel (lower). The statistical indicators, correlation/regression indicators, dynamic query, and discrete aerial perspective line shading features are illustrated on the ASA cars data set [4]. System examples with hurricane trend data are shown in the remainder of this paper.

2 Related Work

The parallel coordinates visualization technique was first introduced by Inselberg [6, 7] to represent hyper-dimensional geometries. Later, Wegman [8] applied the technique to the analysis of multivariate relationships in data. Since then, several innovative extensions to the technique have been described in the visualization research literature.

The system described in this paper implements a histogram display, dynamic axis re-ordering capability, axis inversion, and some details-on-demand features similar to those described by Hauser et al. [9]. In addition, some interaction capabilities described by Siirtola [10] (e.g., conjunctive queries) are included, as well as a variant of the interactive aerial perspective shading technique described by Jankun-Kelly and Waters [11]. The system also includes a focus+context technique for axis scaling that is similar to the capabilities described by Fua et al. [12], Artero et al. [13], Johansson et al. [14], and Novotńy and Hauser [15]. More recently, the coupling of PCP, scatterplots, and correlation computations described by Qu et al. [16] inspired the correlation analysis capabilities in the system described in this paper.

The system also provides dynamic query capabilities based on the double slider concept of Ahlberg and Shneiderman [17]. The PCP axes display important frequency information between the double sliders in a manner similar to the Influence Explorer described by Tweedie [18].

The visualization system described in this paper provides a unique PCP-based interface by fusing variants of the above-mentioned interaction techniques. Another novel contribution from this work is the coupling of this system with statistical indicators and automated analyses. is another novel contribution from this work. This research also describes one of the most in-depth validations of enhanced PCPs in the weather science domain.

Multiple regression traditionally has been used to identify statistically significant variables from multivariate data sets, including tropical cyclone data sets. Klotzbach et al. [19] used this technique to determine the most important variables for predicting the frequency of North Atlantic tropical cyclone activity. Similarly, Fitzpatrick [20] applied stepwise regression analysis to the prediction of tropical cyclone intensity. It will be shown that multiple regression and interactive PCPs can compliment each other, with the regression identifying the relevant associations and the PCPs highlighting additional features of the variables.

3 System Overview

This research has resulted in the development of an innovative visualization system that combines interactive PCP techniques with automated statistical processes to provide a practical tool for analyzing multivariate data sets. The system was developed using the Java Development Kit (JDK) version 1.5; and it yields interactive performance on a laptop computer with a 2.33 GHz Intel Core 2 Duo processor, 3 GB Random Access Memory (RAM), and an ATI Radeon X1600 graphics card with 256 MB Video RAM.

As shown in Fig. 1, the system provides an efficient graphical user interface (GUI) that offers a settings panel (upper left panel), an interactive table view of axis settings and statistics (lower panel), and an enhanced PCP view (upper right panel). Although the table and settings panels are important for the usability of the system, the PCP panel is the heart of the system's visual analysis capabilities. In this panel, the classical PCP method is extended with dynamic interaction capabilities that provide access to the data behind the visualization. The PCP view is dynamically linked with statistical indicators and automatic statistical processes to provide an ideal environment for exploratory data analysis. In the remainder of this section, the principal visualization and statistical analysis capabilities of the system are described.

3.1 Visualization Capabilities

The visualization capabilities of the system are contained in the PCP panel. In addition to many fundamental PCP capabilities such as relocatable axes, axis inversion, and details-on-demand, this panel provides several innovative and intuitive interaction capabilities such as axis scaling (focus+context), aerial perspective shading, and dynamic visual queries. In this subsection, the most significant visualization capabilities of the PCP panel are highlighted.

3.1.1 Dynamic Visual Queries

Since the viewer is often interested in grouping subsets of data, a method to dynamically select lines is provided for each axis. As shown in Fig. 2, each axis has a pair of sliders (the large black triangles) which define the top and bottom range for the query area. Using the mouse cursor, the viewer can drag these sliders to dynamically highlight different lines. Lines within the query area of every axis are rendered with a more prominent, dark color while the remaining lines are rendered with a less prominent, lighter shade of gray. In Fig. 1, an example of a dynamically created conjunctive query is shown using the popular American Statistical Association (ASA) cars data set [4]. In this figure, car records from more recent years (selected on the *Year* axis) are highlighted across the plot.

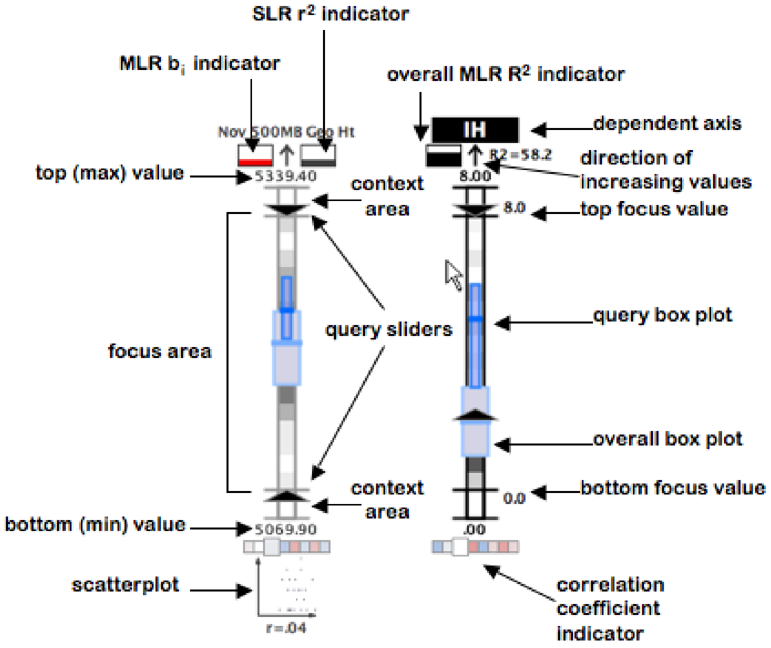


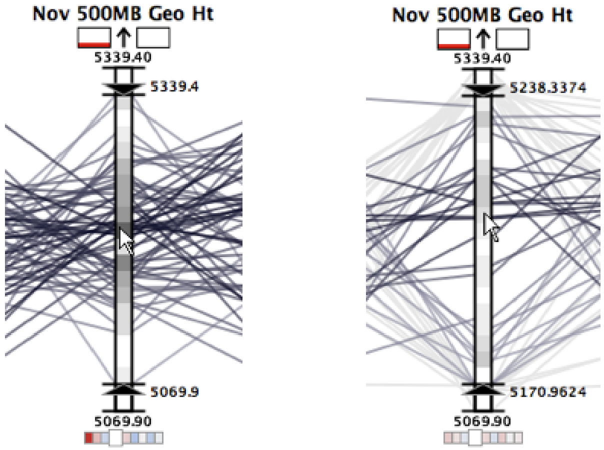
Figure 2: An annotated view of the PCP axis display widget for the system highlighting the visual interaction components and statistical indicators. The axis shown on the left illustrates the normal axis shading while the axis on the right illustrates a highlighted, dependent axis shading.

3.1.2 Axis Scaling (Focus+Context)

The system's dynamic axis scaling capability provides a method to interactively tunnel through the data until a smaller subset of the original data is in focus. Our application allows the user to modify the minimum and maximum focus area values for a selected axis using mouse wheel movement. As shown in Fig. 2, each axis is partitioned into three sections delineated by horizontal tick marks: the central focus area and the top and bottom context areas. When the mouse is hovering over the focus area, an upward mouse wheel motion expands the display of the focus area outward and pushes outliers into the context areas. A downward mouse wheel motion causes the inverse effect: focus region compression. Alternatively, the user may use the mouse wheel over either of the two context areas to alter the minimum or maximum values separately. The user may also manually enter the minimum and maximum values by typing them in appropriate fields of the table view panel. As illustrated in Fig. 3, this intuitive axis scaling capability helps to free space and reduce line clutter, thereby making it easier to analyze relation lines of interest.

3.1.3 Aerial Perspective

The system offers an innovative line shading scheme that is useful for rapidly monitoring trends due to the similarity of data values over multiple dimensions. This shading scheme simulates the human perception of aerial perspective, whereby objects in the distance appear faded while objects nearer to the eye seem more vivid. In this implementation, aerial perspective shading can be used in either a



(a) Before Scaling (b) After Scaling

Figure 3: A screen shot of the parallel coordinates application before (a) and after (b) axis scaling has been performed. In this example, scaling occurs by performing an upward mouse wheel function in the focus area of the axis which moves the values for the top and bottom closer together, effectively stretching the display upward and downward (with the base of the display fixed).

discrete or a continuous mode. In the discrete mode, the lines are colored according to the axis region that they intersect. If any point of a relation line is in the context (non-focus) area of at least one axis, the line is shaded with a light gray color and drawn beneath the non-context lines. If all the points on a relation line fall within the query area of each axis (the area between the two query sliders), the line is colored using a dark gray value that attracts the viewer's attention and the remaining lines (non-query and non-context) are colored a shade of gray that is slightly darker than the context lines but lighter than the query lines. The resulting discrete shading effect is illustrated in Fig. 1.

In the continuous mode, non-context lines go through an additional step to encode the distance of the line from the mouse cursor. As shown in Fig. 3 and Fig. 6(a), query lines that are nearest to the mouse cursor receive the darkest value while lines farthest from the mouse cursor are shaded with a lighter gray. The other query lines are shaded according to a non-linear fall-off function that yields a gradient of colors between said extremes. Consequently, the lines that are nearest to the mouse cursor are more prominent to the viewer due to the color and depth ordering treatments and the viewer can effectively use the mouse to quickly interrogate the data set.

3.2 Statistical Analysis Capabilities

3.2.1 Descriptive Statistical Indicators

To support the interactive analysis capabilities of the system, each axis offers visual representations of key descriptive statistics, identified in Fig. 2. The median, interquartile range (IQR), and the frequency information are calculated for the data in the focus area of each axis. Alternatively, the user can configure the system to display the mean and standard deviation range. These central tendency and variability measures provide a numerical value that indicates the typical value and how "spread out" the samples are in the distribution, respectively. The overall box plots represent the descriptive statistics for all the axis samples, while the query box plots capture the descriptive statistics for the samples that are selected with the axis query sliders. In each axis interior, the frequency information is also displayed by representing histogram bins as small rectangles with gray values that are indicative of the number of lines that pass through the bin's region (see Fig. 2). That is, the darkest bins have the most lines passing through while lighter bins have fewer lines. In Fig. 3, the histogram display is illustrated during an axis scaling operation.

3.2.2 Correlation Analysis Indicators

In statistics, correlation analysis attempts to measure the strength of relationships between pairs of variables. The relationship between two variables can be quantified using a single number, r, that is called the correlation coefficient. Specifically, the system uses the Pearson product-moment correlation coefficient (also called the sample correlation coefficient) to measure the correlation for a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \ldots, n$ [21]. r is given by:

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum {x_i}^2 - (\sum x_i)^2][n \sum {y_i}^2 - (\sum y_i)^2]}} \tag{1}$$

There are two directions or types of correlation: positive and negative r. With a positive correlation, as values of one variable increase, values of the other variable also increase. With a negative correlation, as values of one variable increase, the values of the other variable decrease. Both positive and negative correlations range in strength from weak to strong. A value of zero will occur when the sample points show no linear relationship, the weakest correlation. A perfect linear relationship, the strongest correlation, appears in the sample data when $r = \pm 1$, where $+1$ is a perfect positive relationship and -1 is a perfect negative relationship. In practice, r is rarely perfect as it usually lies somewhere between -1 and $+1$ [21].

The system computes r for each pair of axes in the display, which results in a correlation matrix. As shown in Fig. 2, the rows from this correlation matrix are displayed graphically beneath each axis as a series of color-coded blocks. Each block uses color to encode the sample correlation coefficient between the axis directly above it and the axis that corresponds to its position in the set of blocks. For example, the first block in the correlation indicators under each axis in Fig. 1 represents the correlation strength between the axis above it and the first axis, the *Cylinders* axis. When the mouse hovers over an axis in the PCP panel, the axis is highlighted and the correlation coefficient blocks corresponding to it below the other axes are enlarged (see Fig. 2). The blocks are colored blue for negative correlations and red for positive correlations. The stronger the correlation, the more saturated the color so that stronger correlations are more prominent. Moreover, when the absolute value of a correlation coefficient is greater than or equal to the significant correlation threshold, the block is colored with the fully saturated color. The significant correlation threshold is a user defined value that is also displayed at the bottom of the PCP (see Fig. 1).

In addition to the sample correlation coefficient indicators, the system also displays small scatterplots below the correlation indicators for each axis when an axis is highlighted (see Fig. 2). These scatterplots are created by plotting the highlighted axis values along the y axis and the values from the axis directly above the plot along the x axis of the scatterplot. Each scatterplot also shows the numerical r value associated with the pair of axes. The scatterplots provide a visual means to quickly confirm the type of correlation (positive or negative) and the strength of the correlation. It is important to note that the type of correlation is also visually detectable in the line configuration of the PCP plot. Unlike the other correlation indicators, the scatterplot is useful for discovering nonlinear relationships between variables. For example, a nonlinear relationship can be observed in a scatterplot even if the correlation coefficient is zero. In Fig. 1, nonlinear relationships are illustrated in the scatterplots beneath the second, third, and fourth axes.

3.2.3 Statistical Regression Analysis Capabilities

Regression analysis is often employed to identify the most relevant relationships in a particular data set. Such techniques are effective for screening data and providing quantitative associations. In addition to simple linear regression (SLR), the system offers stepwise multiple linear regression (MLR) with a backwards glance which selects the optimum number of the most important variables using a predefined significance level [21]. Stepwise regression can complement multivariate visualization by isolating the significant variables in a quantitative fashion. Our system executes a MATLAB script and captures output from the MATLAB's "regress" and "stepwisefit" utilities that perform simple and stepwise regression, respectively. The MATLAB output stream is then parsed and displayed graphically within the PCP panel.

A normalization procedure is also used in the MLR analysis so that equal comparison between the variables can be done. Denoting σ as the standard deviation of a variable, y as the dependent variable, $\overline{x}$ as the predictor mean, and $\overline{y}$ as the dependent variable mean, a number k of statistically significant predictors are normalized by the following regression:

$$(y-\overline{y})/\sigma_y = \sum_{i=1}^{k} b_i(x_i-\overline{x}_i)/\sigma_i \tag{2}$$

Two advantages of this approach are that the importance of a predictor may be assessed by comparing regression coefficients b_i between different variables, and that the y-intercept becomes zero.

With the MLR analysis, extra steps are taken to ensure the proper selection of variables. The initially chosen variables are examined for multicollinearity using an automatic filter; if any variables are correlated with each other by more than the significant correlation threshold, one is removed. In this way, the chosen variables are truly independent of each other.

As shown in Fig. 2, the system visually encodes b in the PCP panel using the box below the axis label and to the left of the arrow. Like a thermometer, the box is filled from the bottom to the top based on the magnitude of b. The box is colored red if the coefficient is positive and blue if it is negative. The box to the right of the arrow encodes the r^2 output from the SLR process. In addition to the coefficients, b_i, the MLR analysis returns an overall R^2 value which provides a quantitative indication of how well the model captures the variance between the predictors and the dependent variable. The box beneath the dependent variable axis encodes the overall R^2 value from the MLR analysis.

When these boxes are filled with a light gray 'X' (see Fig. 6), the value is not defined (the SLR or MLR process has not been executed) or, in the case of the MLR analysis, the variable was excluded during the selection process. It is also important to note that the axis corresponding to the dependent variable is indicated by light gray text on a dark gray box for its title, the reverse shading of the other axes. The dependent axis shading is illustrated by the *IH* axis in Fig. 2.

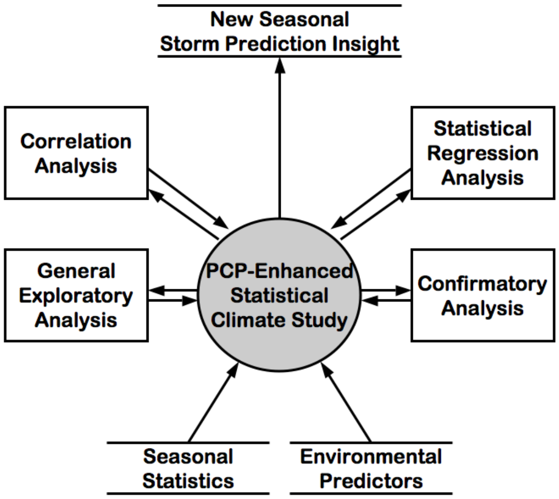

New Seasonal Storm Prediction Insight
Correlation Analysis
Statistical Regression Analysis
General Exploratory Analysis
PCP-Enhanced Statistical Climate Study
Confirmatory Analysis
Seasonal Statistics
Environmental Predictors

Figure 4: System context diagram depicting the workflow for using the PCP system to execute a formal climate study of real-world environmental parameters.

In addition to the multicollinearity filter, the system allows the user to automatically execute the MLR and SLR processes. Furthermore, the system can automatically arrange the axes using the value of b or r^2 from the MLR and SLR analysis, respectively.

4 Enhanced Visual Statistical Analysis Workflow

The visualization capabilities and statistical processes offered by the system provide a unique environment for performing complex multivariate data analysis. During the system development and testing, a systematic workflow was formulated to guide the scientist. In this section, the workflow that is depicted in Fig. 4 will be described. Although this workflow is described in a sequential order, typical analysis involves several iterations and moving between the various processes.

After preparing and loading the data set into the system, the scientist will manually filter the display to remove unnecessary axes. Then, the scientist will manually arrange the variable axes and interact with the display using the previously mentioned visual query techniques. During this initial exploratory analysis, the scientist will acquire a preliminary overview of the entire data set.

Next, the scientist will observe the statistical correlations in the data using the correlation analysis processes and indicators. The system's automated axis arrangement tools can be used in this stage to highlight strong correlations and compare IQR ranges in the data. To prepare for the regression analysis, the scientist can manually reduce multicollinearity by using the correlation indicators to identify and filter correlated variables using a predefined significance level. The scientist can also utilize the automatic multicollinearity filter to ensure that the predictors are truly independent of one another. Removing the strongly correlated independent variables will ultimately improve the MLR analysis by avoiding over-fitting the data. The scientist will gain additional insight in this phase by observing correlations between the predictors as well as correlations between each predictor and the dependent variable.

After the correlation analysis, the scientist will use the integrated SLR processes. This capability provides an alternative indication of the individual associations between the predictors and the dependent variable. The scientist may glean additional insight from this exercise to determine if additional variables should be removed from the view. Then, the scientist is ready to execute the MLR processes in order to quantify the significance of the predictors to the dependent variable. The result of this process is a ranked list of the most important variables for the dependent variable. Unlike the SLR and correlation analysis, the MLR analysis considers the contribution in relation to the other predictors.

By following this workflow, the scientist will develop new ideas about how the specific variables can be used to predict the dependent variable. That is, the scientist will have formed hypotheses about the associations between the variables. Then, the scientist can continue to explore the data in the system to attempt to prove or disprove the new hypotheses; a process that Tukey [22] calls confirmatory data analysis. For example, the scientist may discover patterns in the climate data that will help predict the hurricane activity in 2005 based on the analysis of data from 1950 to 2006. If the theory holds after this testing, the scientist may use the new insight to predict future hurricane activity.

5 Effectiveness Evaluation: A Hurricane Climate Study

The visualization system, concepts, and analysis workflow have been evaluated in a hurricane climate study. The primary objective of this study was to discover the most important predictors for seasonal intense hurricane activity in the North Atlantic to improve forecasting skill. The secondary objective is to identify additional associations between predictors and temporal patterns in the data. In the remainder of this section, the environmental data set and evaluation results are described.

5.1 Climate Study Data

In this climate study, a data set that contains potential environmental predictors observed annually from 1950 to 2006 (57 records) has been analyzed. Table 1 lists the 16 potential environmental predictors from this data set along with their geographical region. This data set was provided by Phil Klotzbach [23] of the Tropical Meteorology Project at Colorado State University, and it is used to predict North Atlantic tropical cyclone activity for the upcoming storm season by categories. Although many categories are considered in practice, the focus of this study is on the number of intense hurricanes (IH) in a hurricane season. A hurricane is classified as intense when its sustained low-level winds are at least 111 mph [24]. Although intense hurricanes account for just over 20% of the tropical storms and hurricanes that strike the United States, these storms warrant special attention because they are responsible for over 80% of the damage [25].

These variables have known relationships to Atlantic tropical cyclone activity. For example, Chu [26] describes how the North Atlantic basin has fewer tropical cyclones during El Niño Southern Oscillation (ENSO) years, and active seasons in La Niña years. Because of this relationship, scientists use ENSO signals as some predictors of seasonal storm activity. In Table 1, variables 1 through 8 are believed to characterize ENSO events.

5.2 Initial Insight (Overview)

After loading the predictors and seasonal storm statistics, the visual analysis tools are used to explore the data set and rearrange the axes. A portion of this initial view is shown in Fig. 5. The first notable

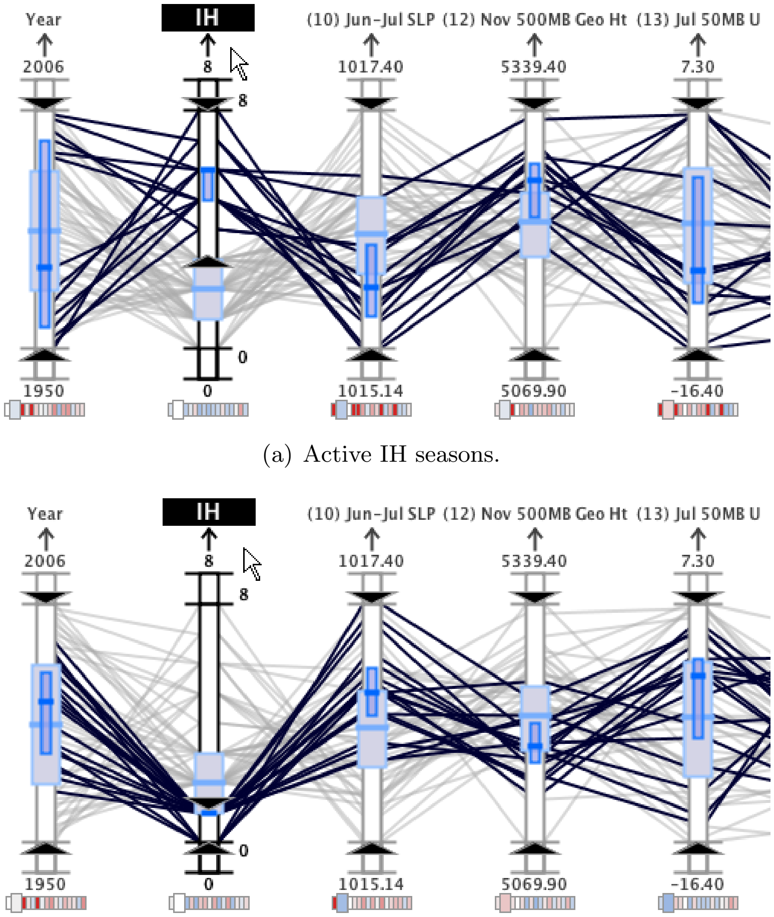
Table 1: Tropical cyclone climate variables evaluated as predictors in the climate study.

	Variable Name	Geographical Region
(1)	June–July Niño 3	5S-5N, 90-150W
(2)	May SST	5S-5N, 90-150W
(3)	February 200-mb U	5S-10N, 35-55W
(4)	February–March 200-mb V	35-62.5S, 70-95E
(5)	February SLP	0-45S, 90-180W
(6)	October–November SLP	45-60N, 120-160W
(7)	Sept. 500-mb Geopotential Height	35-55N, 100-120W
(8)	November SLP	7.5-22.5N, 125-175W
(9)	March–April SLP	0-20N, 0-40W
(10)	June–July SLP	10-25N, 10-60W
(11)	September–November SLP	15-35N, 75-97W
(12)	Nov. 500-mb Geopotential Height	67.5-85N, 50W-10E
(13)	July 50-mb U	5S-5N, 0-360
(14)	February SST	35-50N, 10-30W
(15)	April–May SST	30-45N, 10-30W
(16)	June–July SST	20-40N, 15-35W

SST – Sea Surface Temperature
SLP – Sea Level Pressure

observation is that most of the predictors have low variability (evident by the relatively small overall IQRs) except for the *July 50 mb Equatorial Wind (U) around the globe* (13) predictor (the last axis in Fig. 5). Since the objective is to use the climate variables to predict inactive or active seasons, the overall axis box plot is used to identify the seasons with normal IH activity. That is, the seasons that cross the axis within the box plot are considered normal. Then, the query sliders are used to investigate the behavior of each axis in active (above normal) and inactive (below normal) seasons. In Fig. 5, the active (a) and inactive (b) IH seasons are highlighted. Focusing on the narrower query box plots reveals that some variables, such as *June–July SLP in the tropical Atlantic* (10) and *November 500 mb Geopotential Height in the far North Atlantic* (12), exhibit significantly different behavior in active versus inactive seasons. That is, in active years, the values for (10) are low and the values for (12) are high whereas the opposite conditions are observed in inactive years.

In addition, a gap is visible on the *Year* axis (the first axis in Fig. 5 (a)) for the active seasons. From 1960 to 1994, a relatively quiet period is observed since there are no seasons with an above normal number of intense hurricanes. What's more, Fig. 5 (b) shows that the inactive seasons are clustered into this same time of normal or below normal activity. This visual observation agrees with findings published in the weather research literature [19, 25, 27] that suggest a strong multidecadal variability in the number of intense hurricanes per year in the North Atlantic.



(a) Active IH seasons.

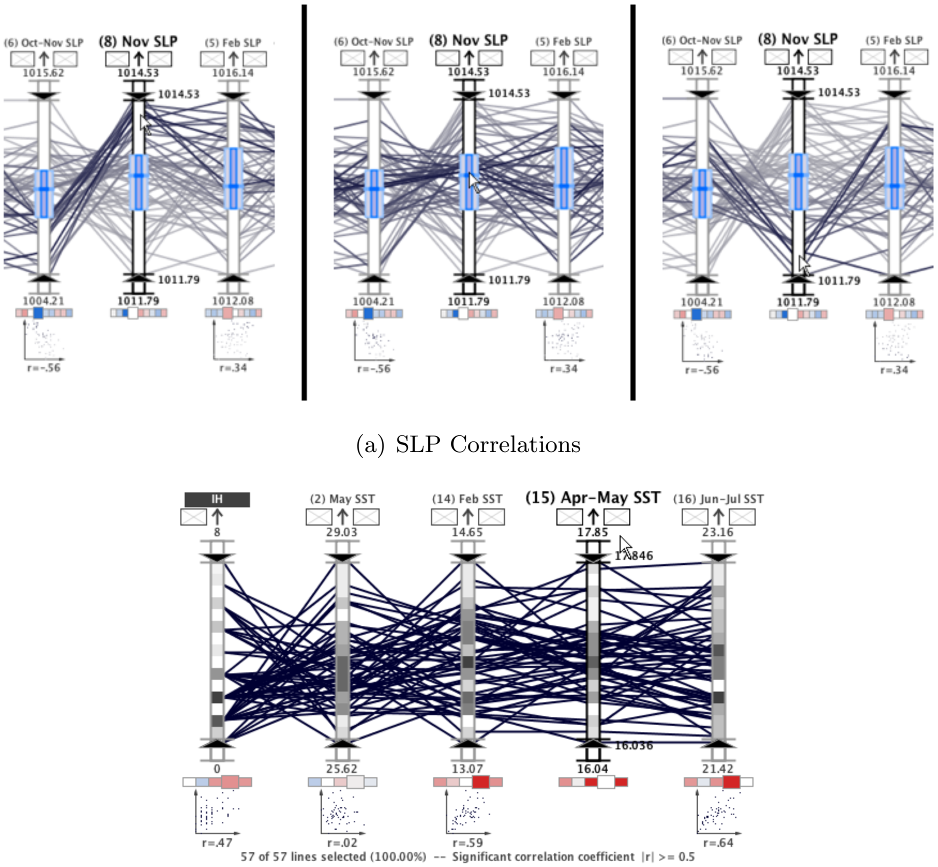
(b) Inactive IH seasons.

Figure 5: A portion of the initial PCP of the intense hurricane seasons partitioned by activity. The active seasons are highlighted in (a) while the inactive seasons are highlighted in (b). From 1960 to 1994, a gap in the seasons with above normal intense hurricane activity is revealed in (a) and the below normal seasons fill this gap in (b).

5.3 Correlation Analysis

To prepare for the MLR analysis and to address the secondary objective of the study, the correlations between the axes are investigated by arranging the 16 axes by the correlation coefficient with the *IH* axis. The correlation indicators reveal that the strongest correlations with the *IH* axis are *June–July SLP in the tropical Atlantic* (10) and *November 500 mb Geopotential Height in the far North Atlantic* (12) — the axes directly to the left and right of *IH* in Fig. 7, respectively. More specifically, the enlarged color-coded correlation indicator box, PCP polyline 'X'-shaped crossings, downward slope in the scatterplot, and numerical display of r in this plot reveal that axis (10) has the strongest negative correlation. Likewise, the strongest positive correlation with axis (12) is evident by the correlation indicator, the more parallel PCP polyline configuration, the upward slope of the scatterplot, and the numerical display of r.

The image sequence shown in Fig. 6(a) illustrates the use of the continuous aerial perspective



(a) SLP Correlations

(b) SST Correlations

Figure 6: Correlation analysis can be performed rapidly using the shading and statistical indicators. In (a) a sequence of images demonstrates how the aerial perspective shading can be used to analyze the SLP variable correlations by moving the mouse from the top to the bottom of the axis. In (b), the correlations between the 4 SST variables are examined revealing the strong positive correlation of variable (15) with both (14) and (16).

shading capability to investigate a strong negative correlation between *October–November SLP in the Gulf of Alaska* (6) and *November SLP in the Subtropical NE Pacific* (8) axes. This intuitive visual query technique, which shades the polylines according to their proximity to the mouse cursor, highlights the 'X'-shaped polyline crossings between the axes, which is indicative of a negative correlation in a PCP.

In Fig. 6(b), the correlations between three SST variables and the *April–May SST off the Northwestern European Coast* (15) variable are shown. In the PCP, strong correlations are identified when $|r| \geq 0.5$, the significant correlation threshold, and visually by a fully saturated correlation indicator. This plot reveals that a relatively strong positive correlation exists between axis (15) and both the *February SST off the Northwestern European Coast* (14) and the *June–July SST in the Northeastern Subtropical Atlantic* (16) axis. Meanwhile, the *May SST in the eastern equatorial Pacific* (2) variable exhibits almost no correlation ($r = .02$). To reduce the multicollinearity between the SST predictors, axis (14) and (16) must be removed since they have a strong correlation with axis (15) and axis (15) has a stronger correlation with the *IH* axis (see Fig. 7). Removing these and any other variables with

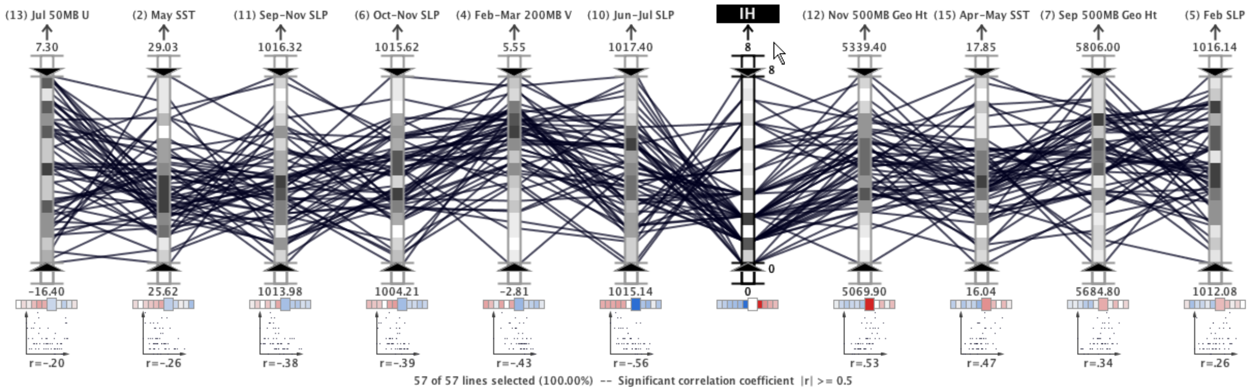


Figure 7: The resulting axis arrangements after applying the multicollinearity filter to ensure independence between the predictors. The axes have been automatically arranged according to the strength of the correlation between the predictors and the dependent axis. Negatively correlated axes are placed to the right of the dependent axis, *IH*, while positively correlated axes are placed to the left.

strong correlations between predictors will ensure the independence of the predictors and thus improve the MLR analysis results.

Before removing axis (14) and (16), the physical relationships between these variables can be considered in order to investigate the cause of the strong correlation. From the geographic extents of these variables listed in Table 1, one can observe that the 3 predictors with strong correlations are all sampled in the North Atlantic Ocean. However, axis (2), which exhibits a very weak correlation, is measured in the Pacific Ocean. Therefore, the strong correlations among axis (14), (15) and (16) can be mostly attributed to the close geographical proximity of the measurements whereas the low correlation of axis (2) can be attributed to the fact that it is measured in the Pacific ocean.

The scientist can continue to manually find and eliminate the highly correlated predictors, or use the system's automatic multicollinearity filter. Applying this filter to the climate data set removes *March–April SLP in the eastern tropical Atlantic* (9) (because of its strong correlation with axis (10)), axis (14) and (16) (strong correlation with axis (15)), *November SLP in the Subtropical NE Pacific* (8) (strong correlation with *October–November SLP in the Gulf of Alaska* (6)), *June–July Niño 3* (1) (strong correlation with axis (2)), and *February 200 mb zonal wind (U) in Equatorial East Brazil* (3) (strong correlation with *February SLP in the Southeast Pacific* (5)). In Fig. 7, the resulting axis configuration is shown, automatically arranged by the correlation coefficient with the *IH* axis. In this plot, it is clear that the only remaining r values greater than the significant correlation threshold (visually indicated by the fully saturated fill color in the enlarged correlation indicators) are the two axes on either side of the *IH* axis; but these correlations are with the dependent axis which does not affect the independence between the predictors.

5.4 Identifying Most Important Predictors

Using the system's automatic SLR and stepwise MLR processes, the predictors are automatically analyzed to determine the most important predictors with respect to the number of intense hurricanes in a season. In Fig. 8, the results of the MLR and SLR analysis are shown. Here the predictors are arranged according to the magnitude of the MLR coefficient, b. The significance level in the stepwise regression analysis was 80%.

Table 2: Significant climate variables for number of intense hurricanes in 1950–2006.

Number of Intense Hurricanes (IH)
(R^2 is 58% and Adjusted R^2 is 54%)

Chosen Variables	Normalized Coefficients b	Sample Mean
Nov. 500-mb Geopot. Ht. (12)	0.3524	5213.38
June–July SLP (10)	–0.3121	1016.23
Sep. 500–mb Geopot. Ht. (7)	0.2514	5753.33
Feb.–Mar. 200-mb V (4)	–0.1871	2.53
Sep.–Nov. SLP (11)	–0.1431	1014.98

The numerical results of the regression listed in Table 2 and the visual representation in Fig. 8 suggest that the five chosen variables are the most significant predictors for the number of intense hurricanes in a season. Highlighting the active and inactive ranges in Fig. 8 also reveals how each specific variable behaves in either active or inactive seasons.

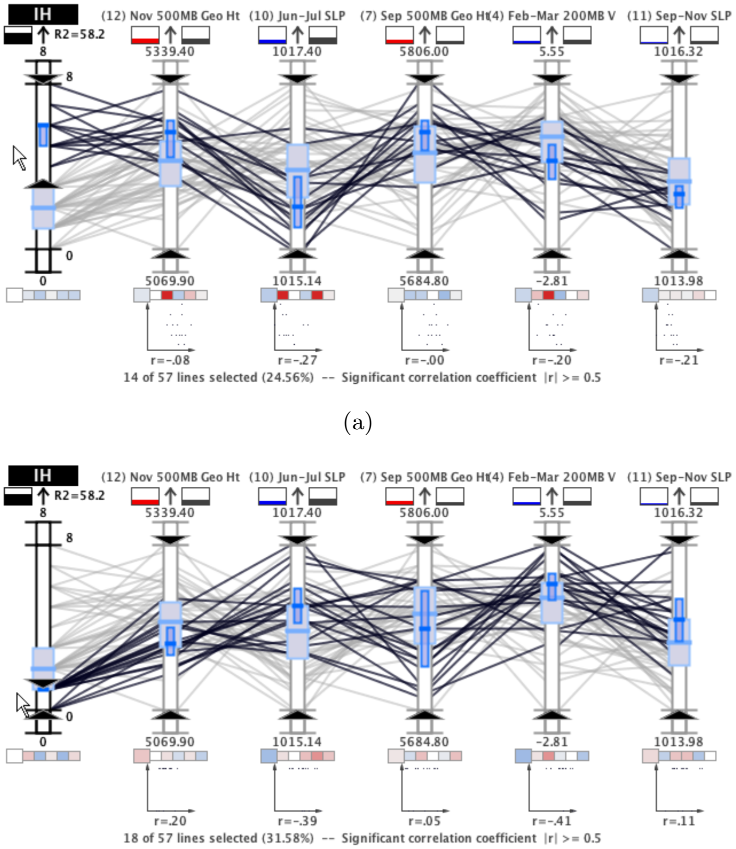
In Fig. 9, the query sliders are used to highlight the points with high values on axis (12), low values for axis (16), low values for axis (7), high values for axis (4), and low values for axis (11). This plot reveals that using these axis ranges to predict the intense hurricanes of a season would result in successfully identifying 11 of the 14 seasons (74%) that had a high number of intense hurricanes between 1950 and 2006. On the other hand, using this technique might result in missing 3 above normal activity seasons (with 7, 6, and 5 intense hurricanes). In particular, one of the storm seasons that is not selected by this query is the infamous 2005 hurricane season which had 7 intense hurricanes, including the cataclysmic Hurricane Katrina. Using the visual query capabilities, minor adjustments can be applied to the query sliders of the significant predictors to ensure that all 14 seasons with active intense hurricane activity are captured. Then, these numerical predictor ranges can be used to predict the activity of future tropical cyclone seasons with respect to the number of intense hurricanes.

5.5 Confirmatory Analysis

To be complete, the physical relationships of the selected predictors can be evaluated to ascertain the validity of the selections from a weather science perspective. Although a detailed physical evaluation is beyond the scope of this article, the selections of these five predictors can be validated by briefly describing how each variable influences the development of tropical cyclones.

The most significant predictor, axis (12), measures the the long-term oscillations which impact global wind patterns, known as teleconnections. When these oscillations are in one phase, they cause more ridges in the Atlantic, which corresponds to less wind shear. Also, weaker zonal winds in the subpolar areas are indicative of a relatively strong thermohaline circulation and therefore a warmer Atlantic Ocean. The MLR results indicate that when predictor (12) is normal or above normal, the environment is more favorable for the development of intense hurricane systems.

Pressure in the Atlantic Ocean is inversely related to tropical cyclone activity; low sea-level pressure in the tropical Atlantic implies increased atmospheric instability, moisture, and ascent (more favorable for the genesis of tropical cyclones), and weaker trade winds (which correspond to less wind shear that



(a)

(b)

Figure 8: Results of the MLR analysis showing the axes arranged in descending order based on the MLR result coefficient, *b*. In (a) the active seasons are highlighted and in (b) the inactive seasons are highlighted. The query box plots in (a) are always entirely above or below the overall median for each axis which reinforces the predictability of these variables with respect to intense hurricane activity.

can tear up the thunderstorms in tropical cyclones). This relationship explains the selection of axis (11) and axis (10), which are normal or below normal in the active intense hurricane seasons.

The MLR analysis also identified two variables that characterize El Niño events which inhibit tropical cyclone formation and intensification in the Atlantic. The first clues of an impending El Niño can be detected in February by observing three variables. The MLR analysis selected one of these variables, axis (4), which measures the anomalous late winter north-south winds at 200 mb in the southern Indian Ocean (a condition associated with El Niño). As shown in Fig. 9, normal to below normal values of (4) correspond to more favorable conditions for intense hurricane development. The MLR model includes one Fall variable that is correlated to El Niño conditions for the following year, axis (7), which is more favorable for hurricane intensification in normal to above normal measurements.

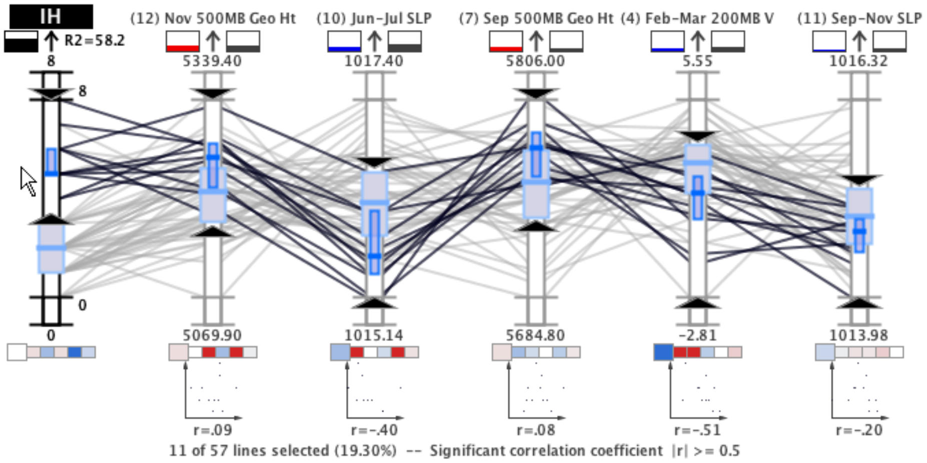


Figure 9: The query sliders and the MLR results are used to highlight the ranges of the most important predictors. The dynamic query capabilities of the system are exploited to interactively confirm the theory that these predictor ranges can be used to forecast intense hurricane activity.

6 Conclusion

This research has demonstrated that interactive parallel coordinates, a visualization technique designed specifically for complex multivariate information, can be used in conjunction with advanced statistical analysis to discover and confirm hypotheses. While the regression analysis yields an ordering of the most important predictors, the dynamic visual analysis capabilities of the system facilitate a deeper understanding of the associations. Using traditional analysis alone would require the examination of 136 scatterplots to observe the same associations in the data that are efficiently captured by the interactive visualization system presented in this paper.

During the development and evaluation of the visualization system, a systematic workflow for analyzing complex climate study data has been formulated. Using this workflow, the effectiveness of the concepts that emerged in this research are demonstrated in a real-world case study to identify the most significant predictors for the number of intense hurricanes in a hurricane season. In the future, these results will be expanded to include additional seasonal statistics and climate study data sets. In addition, new multivariate visualization capabilities will be developed that enhance the study of climate data, thus giving researchers a more effective visual alternative for understanding the climate.

Acknowledgements

This research is sponsored by the Naval Research Laboratory's Long-Term Training Program and the National Oceanographic and Atmospheric Administration (NOAA) with grants NA060AR4600181 and NA050AR4601145, and through the Northern Gulf Institute funded by grant NA06OAR4320264. The authors wish to thank Dr. Phil Klotzbach of Colorado State University's Tropical Meteorology Project for providing the Atlantic tropical cyclone data set.

References

[1] F. Vitart, "Dynamical seasonal forecasts of tropical storm statistics," in *Hurricanes and Typhoons: Past, Present, and Future* (R. J. Murnane and K.-B. Liu, eds.), pp. 354–392, Columbia University Press, Dec. 2004.

[2] R. A. Rensink, "Change detection," *Annual Review of Psychology*, vol. 53, pp. 245–577, 2002.

[3] C. G. Healey, L. Tateosian, J. T. Enns, and M. Remple, "Perceptually-based brush strokes for nonphotorealistic visualization," *ACM Transactions on Graphics*, vol. 23, no. 1, pp. 64–96, 2004.

[4] "The cars data set." http://stat.cmu.edu/datasets (current 16 Jan. 2008).

[5] C. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, eds., *NIH/NSF Visualization Reserach Challenges.* IEEE Press, 2006. http://tab.computer.org/vgtc/vrc/index.html (current 31 Mar. 2008).

[6] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 4, pp. 69–91, 1985.

[7] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proceedings of IEEE Visualization 1990*, (San Francisco, CA), pp. 361–378, IEEE Computer Society, 1990.

[8] E. J. Wegman, "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 664–675, 1990.

[9] H. Hauser, F. Ledermann, and H. Doleisch, "Angular brushing of extended parallel coordinates," in *Proceedings of IEEE Symposium on Information Visualization 2002*, (Boston, MA), pp. 127–130, IEEE Computer Society, 2002.

[10] H. Siirtola, "Direct manipulation of parallel coordinates," in *Proceedings of the International Conference on Information Visualisation*, (London, England), pp. 373–378, IEEE Computer Society, 2000.

[11] T. J. Jankun-Kelly and C. Waters, "Illustrative rendering for information visualization," in *Posters Compendium: IEEE Visualization 2006*, (Baltimore, MD), pp. 42–43, IEEE Computer Society, 2006.

[12] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," in *Proceedings of IEEE Visualization*, (San Francisco, California), pp. 43–50, IEEE Computer Society, Oct. 1999.

[13] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz, "Uncovering clusters in crowded parallel coordinates visualization," in *IEEE Symposium on Information Visualization*, (Austin, Texas), pp. 81–88, IEEE Computer Society, Oct. 2004.

[14] J. Johansson, P. Ljung, M. Jern, and M. Cooper, "Revealing structure within clustered parallel coordinates displays," in *IEEE Symposium on Information Visualization*, (Minneapolis, Minnesota), pp. 125–132, IEEE Computer Society, Oct. 2005.

[15] M. Novotńy and H. Hauser, "Outlier-preserving focus+context visualization in parallel coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 893–900, 2006.

[16] H. Qu, W. Chan, A. Xu, K. Chung, K. Lau, and P. Guo, "Visual analysis of the air pollution problem in hong kong," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, pp. 1408–1415, November/December 2007.

[17] C. Ahlberg and B. Shneiderman, "Visual information seeking: Tight coupling of dynamic query filters with starfield displays," in *Proceedings of Human Factors in Computing Systems*, (Boston, MA), pp. 313–317, 479–480, ACM, 1994.

[18] L. Tweedie, R. Spence, H. Dawkes, and H. Su, "Externalising abstract mathematical models," in *Proceedings of the Conference on Human Factors in Computing Systems*, (Vancouver, British Columbia, Canada), pp. 406–412, ACM, 1996.

[19] P. J. Klotzbach, W. M. Gray, and W. Thorson, "Extended range forecast of Atlantic seasonal hurricane activity and U.S. landfall strike probability for 2007," tech. rep., 2006. http://tropical.atmos.colostate.edu/Forecasts/2006/dec2006/ (current 31 Mar. 2008).

[20] P. J. Fitzpatrick, *Understanding and Forecasting Tropical Cyclone Intensity Change*. PhD thesis, Department of Atmospheric Sciences, Colorado State University, Fort Collins, CO, 1996.

[21] R. E. Walpole and R. H. Myers, *Probability and Statistics for Engineers and Scientists*. Englewood Cliffs, New Jersey: Prentice Hall, 5th ed., 1993.

[22] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[23] P. J. Klotzbach. personal communication, Jan. 2007.

[24] P. J. Fitzpatrick, *Natural Disasters, Hurricanes: A Reference Handbook*. Santa Barbara, California: ABC–CLIO, 1999.

[25] S. B. Goldenberg, C. W. Landsea, A. M. Mestas-Nuñez, and W. M. Gray, "The recent increase in atlantic hurricane activity: Causes and implications," *Science*, vol. 293, pp. 474–479, July 2001.

[26] P.-S. Chu, "ENSO and tropical cyclone activity," in *Hurricanes and Typhoons: Past, Present, and Future* (R. J. Murnane and K.-B. Liu, eds.), pp. 297–332, Columbia University Press, 2004.

[27] P. J. Klotzbach and W. M. Gray, "Summary of 2006 atlantic tropical cyclone activity and verification of author's seasonal and monthly forecasts," tech. rep., Nov. 2006. http://hurricane.atmos.colostate.edu/Forecasts/2006/nov2006/ (current 31 Mar. 2008).